# RELATÓRIO DE ATIVIDADES N. 2

Produto 2 - Contrato de serviço CON22-00016732
Contratado: Krerley Irraciel Martins Oliveira

RELATÓRIO DE ATIVIDADES RELATIVO AO PRODUTO 2 DESCRITO NO TERMO DE REFERÊNCIA - PROJETO OMS “PLATAFORMA CLÍNICA GLOBAL COVID-19 E PÓS-COVID” - CONTRATO POR SERVIÇOS – CON22-00016732.

MACEIÓ, 28 DE NOVEMBRO DE 2022

# Descrição do produto de acordo com Termo de Referência:

**Produto 2 -** Produto 2 - Relatório técnico contendo diagnóstico das limitações associadas ao processo de mineração de dados, coleta de dados anonimizados dos pacientes com sequelas de médio e longo prazo covid-19 e propostas de estratégias complementares para qualificação da alimentação da Plataforma Clínica Global – OMS, período de janeiro/2021 a maio/22, referente à Estratégia 2: Seguimento dos Pacientes Pós-Covid), referente à pesquisa do Projeto OMS “Plataforma Clínica Global Covid-19 e Pós-Covid”, com estudo observacional (retrospectivo), do seguimento dos pacientes com sequelas de médio e longo prazo covid-19, a continuidade do cuidado no pós-covid referente ao atendimento de suas necessidades clínicas e de reabilitação, no âmbito da Rede de Atenção à Saúde da Secretaria Municipal de Saúde de Florianópolis.
CONTRATO DE SERVIÇO CON22-00016732, KRERLEY IRRACIEL MARTINS OLIVEIRA

# Lista de arquivos que acompanham o relatório:

(disponíveis no link: https://github.com/goedert/LED_Floripa_Covid_OMS-Resultados.)

**Projeto:** *PLATAFORMA CLÍNICA GLOBAL COVID-19 E PÓS-COVID*
**Relatório:** Final
**Responsável:** Prof. Krerley Oliveira (UFAL)
**Colaboradores:** Prof. Sergio Lira (UFAL), Guilherme T. Goedert (EMAp/FGV & Univ. Roma Tor Vergata & RWTH Aachen Uni.), Juliano Genari de Araújo (ICMC/USP), Adriano Barbosa (UFGD), Julia Bonfim (CESMAC-AL), Pedro Coutinho (FAMED/UFAL)
**Período de referência:** Novembro/2022

# SUMÁRIO

1 - INTRODUÇÃO;
2 - OBJETIVOS;
3 - METODOLOGIA ;
4 - ANÁLISE DOS DADOS DO ESTUDO
5 - DESENVOLVIMENTO DO ESTUDO DE PROCESSAMENTO DE LINGUAGEM NATURAL
6 - PRINCIPAIS LIMITAÇÕES ENCONTRADAS E PROPOSTAS DE ESTRATÉGIAS COMPLEMENTARES
7- CONCLUSÃO E ALIMENTAÇÃO DA PLATAFORMA GLOBAL
8- REFERÊNCIAS
9- ANEXOS

# 1-INTRODUÇÃO

## 1.1- APRESENTAÇÃO DO TRABALHO

O presente relatório se inclui no esforço global empreendido pela OPAS/OMS para entendimento das sequelas de médio e longo prazo covid-19 e a continuidade do cuidado no pós-covid referente ao atendimento de suas necessidades clínicas e de reabilitação. Tal projeto congrega uma rede global de serviços de assistência à saúde, da qual a Secretaria de Saúde de Florianópolis é parte integrante. Como estratégia metodológica do projeto global, foi decidido desenvolver um estudo observacional retrospectivo. O presente contrato se inclui nesse âmbito e, considerando os parâmetros da OPAS/OMS, tem o escopo de desenvolver estudos para obtenção dos dados referentes às sequelas de médio e longo prazo covid-19, a continuidade do cuidado no pós-covid referente ao atendimento de suas necessidades clínicas e de reabilitação, no âmbito da Rede de Atenção à Saúde da Secretaria Municipal de Saúde de Florianópolis.

O principal objetivo deste trabalho é o desenvolvimento de estudo de mineração e análise de dados anonimizados dos pacientes do município de Florianópolis-SC acometidos por Covid-19 no período de Janeiro/2020 a Setembro/2022 que atendam aos critérios do formulário Case Report Form for COVID-19 sequelae (“Post COVID-19 CRF”), para fins de alimentação da Plataforma Clínica Global-OMS (WHO Global Clinical Platform for COVID-19), bem como a análise de limitações e propostas de estratégias complementares para qualificação da alimentação da Plataforma Clínica Global – OMS.

O projeto tem sido executado em colaboração com a gerência de vigilância epidemiológica da Prefeitura de Florianópolis, com a qual foi acordado desde o princípio a realização de ao menos uma reunião semanal de planejamento e acompanhamento de atividades entre a equipe da vigilância epidemiológica e a equipe do LED/UFAL. Assim, o presente estudo é concomitante e colabora com o estudo realizado sob coordenação do prof. Sérgio Lira, que trata da sistematização da mineração dos dados.

Nas seções a seguir iremos apresentar o trabalho executado ao longo do projeto, que durou de Agosto a Novembro de 2022.

# 1.2- FORMULÁRIO POST COVID-19 CRF

A introdução do Formulário POST COVID-19 CRF inclui resumos executivos sobre a Plataforma Clínica Global para COVID-19 da WHO (descrevendo os formulários de estudos anteriores), bem como os objetivos e critérios referentes ao preenchimento do formulário para acompanhamento de sequelas pós-COVID.

Nota-se que o protocolo estabelecido originalmente pelo estudo é que o formulário seja preenchido em visitas de acompanhamento do paciente em intervalos regulares (formando uma coorte prospectiva): primeiro preenchimento entre 4 e 8 semanas da alta do paciente e preenchimentos subsequentes a cada 3 meses enquanto perdurarem sintomas ou sinais associados a COVID-19 e a cada 6 meses se não houver sintomas ou sinais relevantes.

Foi apontado que este protocolo de inclusão não é compatível com a realidade nacional brasileira, de modo que os centros parceiros que estão realizando este estudo no Brasil adotaram um protocolo retrospectivo a partir da mineração de dados de atendimento de saúde nos respectivos centros. Esta estratégia também está sendo utilizada por nosso grupo no presente projeto.

O questionário que compõe este estudo é formado por **626 perguntas, organizadas em três módulos**: background do participante (demografia, condições pré-existentes e histórico da infecção aguda por COVID-19); histórico de acompanhamento pós-COVID e exames/diagnósticos realizados em retorno após infecção aguda.

A estrutura dos módulos pode ser resumida em:

1. **Módulo 1**: Questões relativas ao background do paciente (**192 perguntas**).
    1. Informação de preenchimento do formulário CORE CRF;
    2. Informações demográficas (Sexo, idade, altura, peso, educação, etinicidade, abuso de susbtâncias, gravidês, etc);
    3. Condições pré-existentes;
    4. Informações do quadro clínico relativo a primeira infecção covid;
    5. Complicações durante quadro grave de covid;
    6. Tratamentos utilizados;
    7. Dados de diagnóstico.

2. **Módulo 2**: Questões relativas ao acompanhamento pós-covid (**101 perguntas**).
    1. Informação de hospitalização após caso grave de covid;
    2. Informação de reinfecção;

3. Status de vacinação;
4. Status de ocupação do paciente;
5. Capacidade do paciente de executar atividades diárias.

3. **Módulo 3**: Questões relativas a exames clínicos e diagnósticos em consulta de retorno (**333 perguntas**).
    1. Exame neurológico;
    2. Exame radiográfico;
    3. Exame sanguíneo;
    4. Testes clínicos;
    5. Novos diagnósticos ou doenças relacionadas ao caso de covid;

# 2 - OBJETIVOS

O objetivo desse estudo é diagnosticar as limitações associadas ao processo de mineração de dados, coleta de dados anonimizados dos pacientes com sequelas de médio e longo prazo covid-19 e elaborar propostas de estratégias complementares para qualificação da alimentação da Plataforma Clínica Global.

## 2.1 - Extração de dados

### CELK

Considerando que o Município de Florianópolis utiliza um moderno sistema de prontuário digital para registro dos atendimentos médicos ambulatoriais, chamado CELK, foi decidido pela equipe de vigilância epidemiológica do município que os dados a serem analisados seriam aqueles provenientes das bases digitais do município, principalmente as bases do CELK.

Assim, a primeira etapa do trabalho consistiu em elaborar um acordo de cooperação entre a prefeitura e o LED/UFAL que permitisse o acesso seguro do ponto de vista jurídico e formal dos pesquisadores às bases de dados da prefeitura. Tal acordo de cooperação foi construído e assinado pelos representantes da UFAL, Prefeitura Municipal de Florianópolis e FUNDEPES, conforme Anexo 3, no fim do mês de Agosto/2022. A liberação do acesso aos dados do CELK para a equipe da UFAL ocorreu no início do mês de Setembro/2022.

O CELK utiliza uma base de dados relacional SQL,  grande e complexa, com mais de 1.000 tabelas e inúmeras relações entre cada tabela.

## Identificação e extração dos dados

Dado a indisponibilidade de dicionário de dados da base de dados do CELK ou base de desenvolvimento com dados falsos, foi utilizado uma base de homologação com dados iguais da base de produção, alimentados com um período de atraso e parcialmente anonimizada (substituído nomes por ids), e uma respectiva interface do sistema que utiliza a mesma base de homologação. As tabelas relevantes foram identificadas por uma exploração manual  da base de dados associada com recomendações da equipe da prefeitura. Para identificação das colunas relevantes e seus valores foram consultados pacientes relevantes realizando o cruzamento dos dados preenchidos entre a interface e a base de dados de homologação.

A extração dos dados foi realizada nas tabelas relevantes com querys SQL que acessam os dados (utilizando as relações entre as tabelas para carregar dados simultaneamente de várias tabelas quando relevante) já realizando uma pré-filtragem, somente carregando dados de pacientes com ficha de investigação de COVID-19 preenchidas e marcadas como caso confirmado. Tais dados foram armazenados de maneira anonimizada (somente guardando a id do paciente) em arquivos CSV em servidor da UFAL para posterior processamento.

## Sobre a utilização da base SIVEP-GRIPE

Inicialmente, previámos a utilização da base SIVEP-GRIPE, referente a internações por síndrome respiratória aguda, como complementação à base de atendimento ambulatorial CELK. No entanto, seria necessário acesso não anonimizado desta base para o cruzamento de pacientes entre as duas bases, e foi decidido juntamente com a Secretaria Municipal de Saúde de Florianópolis que este nível de acesso seria inviável na curta duração do presente projeto. Considerando também que a base de internações constaria apenas por si só não constituiria bons critérios para filtragem de pacientes de interesse no estudo de sequelas de COVID, foi decidido contra a utilização final da versão pública e anonimizada desta base. Esta decisão foi fundamentada no fato da base SIVEP-GRIPE conter apenas casos agudos, sem acompanhamento posterior de sequelas; foi decidido juntamente com a Vigilância Epidemiológica da Secretaria Municipal de Saúde que estes dados não seriam utilizados.

# 3- METODOLOGIA

## 3.1 - Filtragem da base

Da base de dados do CELK, foram extraídos todos os atendimentos referentes a pacientes com diagnóstico de COVID confirmado até 28/09/2022, seja por exame laboratorial ou critério clínico. Nesta etapa, foram identificados 154.944 pacientes, totalizando 3.356.283 atendimentos (média de 19,5 atendimentos por paciente). Por atendimento, entendemos um procedimento registrado na base de dados. Assim, uma visita ao serviço de saúde pode gerar vários atendimentos.

Em seguida, identificamos **pacientes de interesse para estudo Pós-COVID, são estes os pacientes que buscaram ao menos um atendimento na rede municipal de saúde entre 4 e 8 semanas após seu diagnóstico de COVID.** Recuperamos todos os atendimentos destes pacientes de interesse a partir da quarta semana após o diagnóstico.

**Critério de inclusão de pacientes e atendimentos médicos na análise:**

- Pacientes confirmados para covid-19 e com atendimentos na rede de saúde da SMS de Florianópolis de janeiro de 2020 a maio de 2022;
- Pacientes que voltaram ao menos uma vez para uma consulta médica na rede de 4 a 8 semanas após a confirmação para covid-19.
- Foram incluídos todos os atendimentos subsequentes destes pacientes como casos suspeitos de pós-covid.
- Para cada novo atendimento, uma nova ficha CRF Post Covid foi preenchida para o paciente de interesse.

Após a filtragem, nossa base é composta por 28.817 pacientes de interesse, totalizando 492.959 atendimentos (média de 17,1 atendimentos por paciente).

## 3.2- Mapeamento de dados estruturados

Os dados que constavam de forma estruturada nos prontuários eletrônicos do Celk e que puderam preencher campos do formulário CRF Post Covid estão listados abaixo. Para o preenchimento foi utilizado o dicionário de respostas fornecido pela OPAS-OMS (Anexo em Arquivos Complementares do Git).

| **Fontes de dados no CELK** | | **Destino (CRF)** | **Módulo** |
|---|---|---|---|
| **Tabela** | **Coluna** | **Variable / Field Name** | **no CRF** |
| usuario_cadsus | cd_usu_cadsus | participant_id | 1 |
| atendimento | nr_atendimento | record_id | 1 |
| atendimento_primario | peso | height | 1 |
| atendimento_primario | altura | weight | 1 |
| atendimento_primario | gestante | participant_pregnant | 1 |
| atendimento_primario | idade_gestacional | gestational_weeks_unknown | 1 |
| atendimento_primario | idade_gestacional | gestational_age | 1 |
| usuario_cadsus | sg_sexo | sex | 1 |
| usuario_cadsus | nivel_escolaridade | education | 1 |
| vac_aplicacao | ds_vacina | product_name_dose2 | 2 |
| usuario_cadsus | cd_raca | ethnicity | 1 |
| investigacao_agr_covid_19 | doenca_card_cronica | heart_disease | 1 |
| investigacao_agr_covid_19 | doencas_renais_avancado | kidney_disease | 1 |
| investigacao_agr_covid_19 | diabetes | diabetes | 1 |
| investigacao_agr_covid_19 | imunossupressao | immunodeficiency | 1 |
| investigacao_agr_covid_19 | condicao_obesidade | obesity | 1 |
| investigacao_agr_covid_19 | dt_primeiros_sintomas | date_of_onset_of_symptoms | 1 |
| investigacao_agr_covid_19 | data_coleta_teste | diagnosis_of_covid | 1 |
| investigacao_agr_covid | resultado_teste | diagnostic_test | 1 |

| | | | |
|---|---|---|---|
| _19 | | | |
| investigacao_agr_covid_19 | tipo_teste | pcr_test | 1 |
| investigacao_agr_covid_19 | data_coleta_teste | pcr_date | 1 |
| investigacao_agr_covid_19 | tipo_teste | antigen_test | 1 |
| investigacao_agr_covid_19 | data_coleta_teste | if_positive_date_of_positi | 1 |
| investigacao_agr_covid_19 | tipo_teste | antibody_test | 1 |
| investigacao_agr_covid_19 | data_coleta_teste | antibody_test_date | 1 |
| investigacao_agr_covid_19 | obito, internado_util, internado, assintomatico, *outros sintomas* | severity_illness | 1 |
| investigacao_agr_covid_19 | internado_uti, internado, tratamento_domiciliar | level_of_care | 1 |
| investigacao_agr_covid_19 | internado_uti | icu_admission | 1 |
| vac_aplicacao | ds_vacina | number_of_doses | 2 |
| vac_aplicacao | ds_vacina | product_name_dose1 | 2 |
| vac_aplicacao | ds_vacina | specify_other_dose1 | 2 |
| usuario_cadsus | dt_nascimento | age | 1 |
| usuario_cadsus | dt_nascimento | age_type | 1 |
| vac_aplicacao | dt_aplicacao | date_of_vaccine_dose_1 | 2 |
| vac_aplicacao | ds_vacina | specify_other_dose2 | 2 |
| vac_aplicacao | dt_aplicacao | date_of_vaccine_dose_2 | 2 |
| investigacao_agr_covid_19 | data_coleta_teste | reinfection | 2 |

# 4- ANÁLISE DO DADOS DO ESTUDO

## 4.1- Resultados de pacientes com histórico de Covid-19 e atendimento subsequente

Tendo identificado pacientes relevantes ao estudo, coletamos todos os atendimentos possivelmente relevantes, como descrito na Seção 4. O total de pacientes e atendimentos identificados é descrito na Tabela 1.

| | **Pacientes** | **Atendimentos** | **Atendim. por paciente** |
|---|---|---|---|
| Pré-filtragem | 154.944 | 3.356.283 | 19,5 |
| Pós-filtragem | 28.817 | 492.959 | 17,1 |

**Tabela 1.** Número de Pacientes de Interesse

Todos os pacientes acima tiveram COVID confirmado até 28/09/2022. Os atendimentos não são necessariamente relativos a COVID.  são todos os atendimentos registrados para os pacientes com pelo menos uma confirmação de COVID.

Dos 28.817 pacientes de interesse, foram identificados 938 casos de reinfecção, usando o critério que duas confirmações laboratoriais distintas (por teste RT-PCR ou de antígenos) separadas por 90 dias ou mais  indicam casos de reinfecção.

Várias colunas são preenchidas no Celk como checkbox onde não é possível distinguir o valor “não” da falta de preenchimento. Portanto campos como por exemplo comorbidades e internação em UTI apresentam porcentagens baixas de preenchimento pois só contém os valores “sim”.

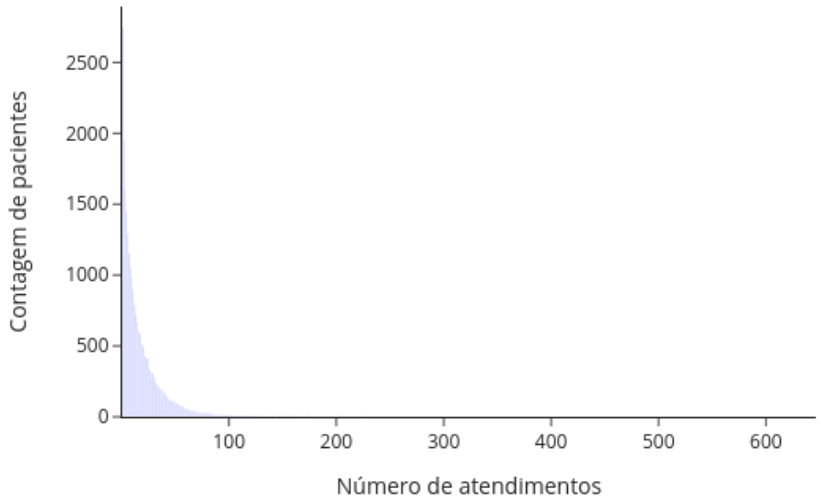



**Figura 1.** Histograma do número de atendimentos para cada paciente antes da filtragem.

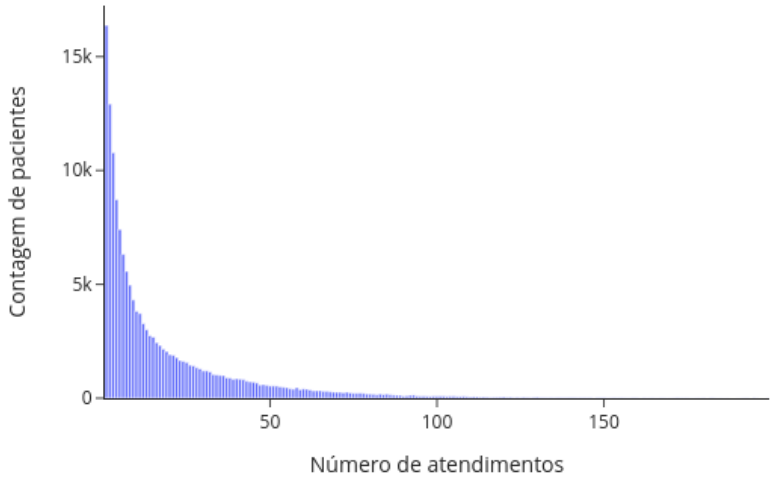



**Figura 2.** Histograma do número de atendimentos para cada paciente depois da filtragem.

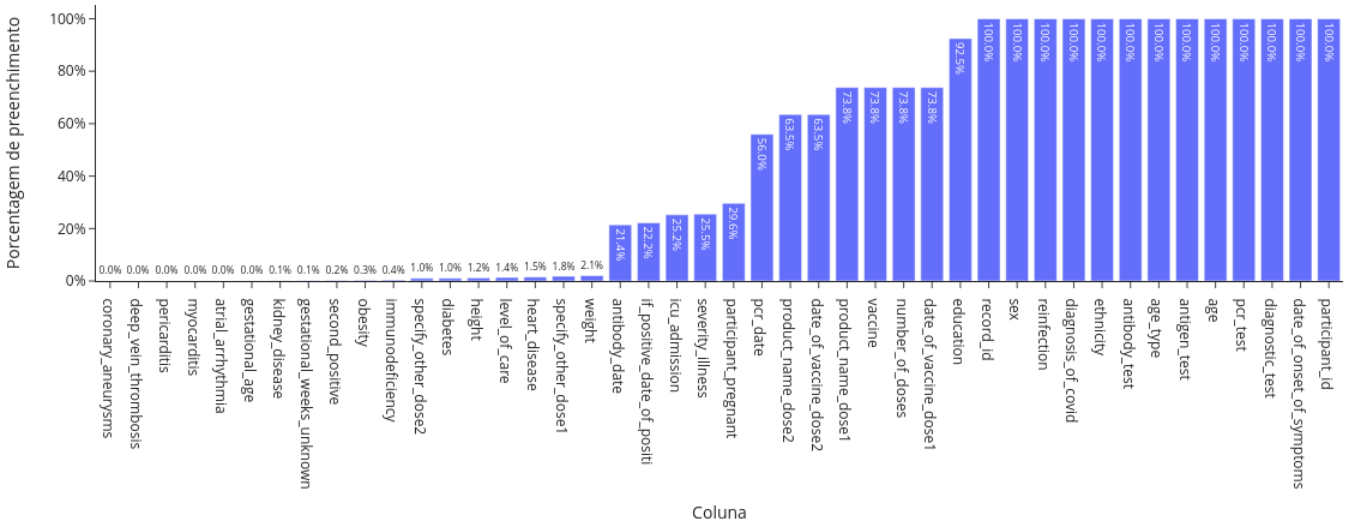



**Figura 3.** Porcentagem de preenchimento dos dados para cada coluna.

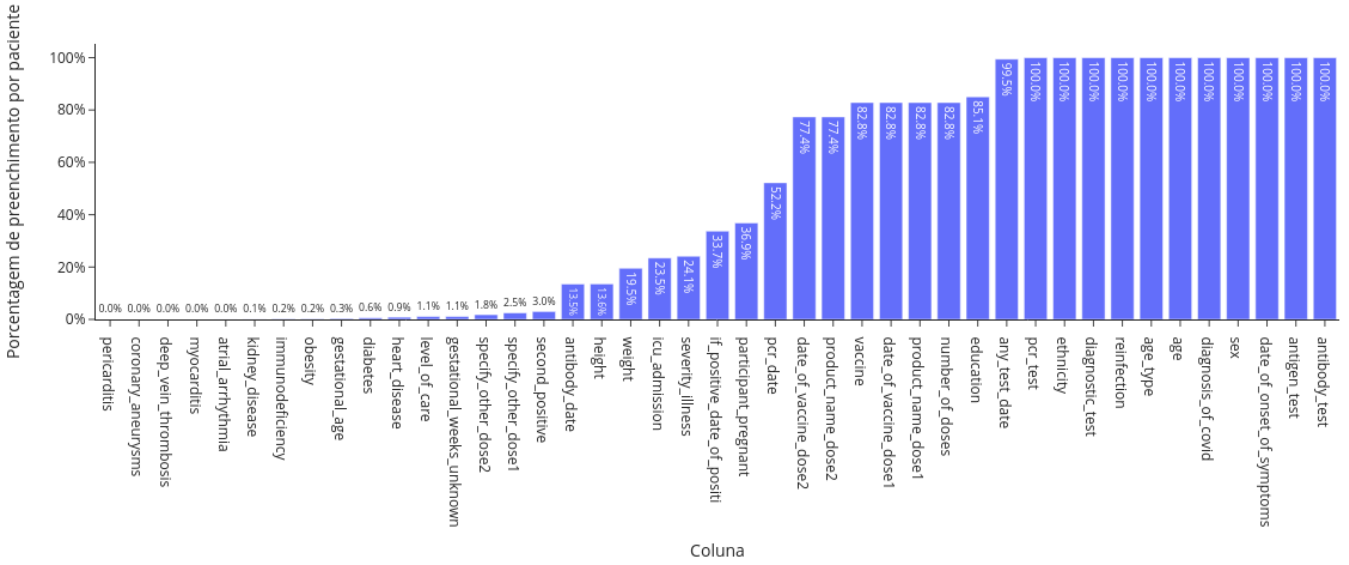



**Figura 4.** Porcentagem de pacientes com pelo menos uma resposta para cada pergunta.

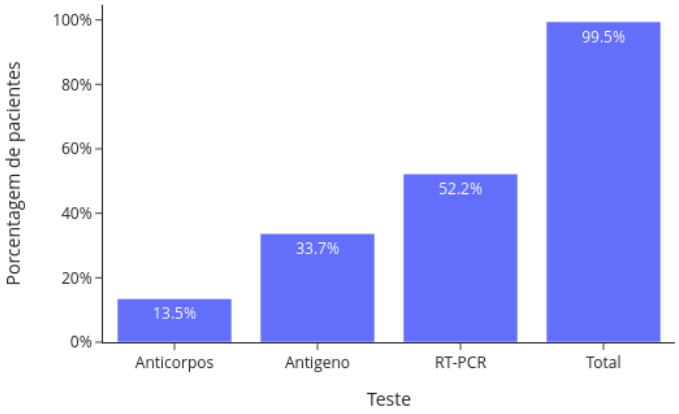



**Figura 5.** Porcentagem de pacientes para os quais é conhecida a data de coleta de teste laboratorial.

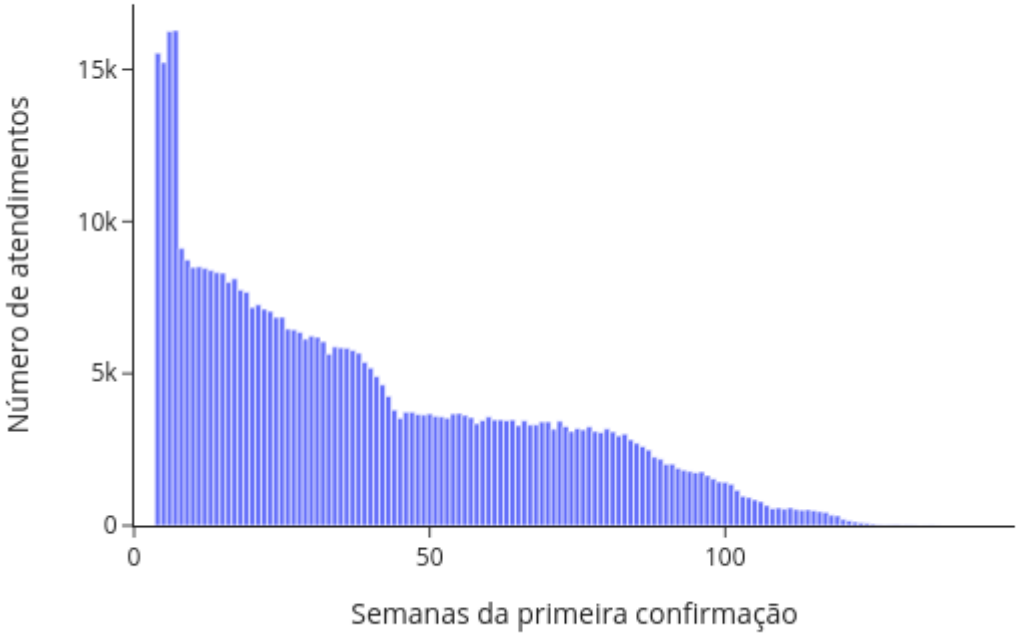



**Figura 6.** Histograma do tempo de retorno entre a primeira confirmação e cada atendimento.

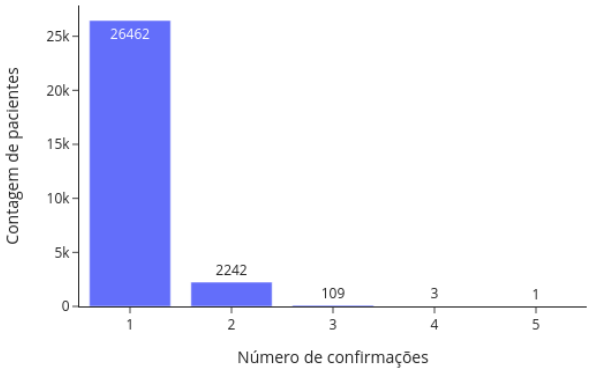



**Figura 7.** Histograma do número de confirmações para os pacientes selecionados.

# 5 - DESENVOLVIMENTO DO ESTUDO DE PROCESSAMENTO DE LINGUAGEM NATURAL

## 5.1 - Plataforma de busca, visualização e rotulação de textos médicos.

Com o objetivo de trabalhar com dados não-estruturados de prontuários médicos presentes na base descrita nas seções anteriores, nosso grupo programou uma plataforma que promove três tarefas em sequência:

- Busca eficiente de textos a partir de uma máquina de busca moderna e flexível, que permite várias estratégias diferentes de busca;
- Representação visual dos textos resultantes na tela através de projeções em pontos, de modo que textos similares estão próximos;
- Rotulação ágil de textos em categorias de interesse, através de uma ferramenta de interface gráfica que permite a manipulação da nuvem de pontos do item anterior.

O uso final desta plataforma permite explorar milhares de textos livres preenchidos pelos profissionais de Saúde de maneira ágil e simples, preparando a informação para a construção de classificadores supervisionados que podem ser utilizados para facilitar o trabalho de preenchimento do relatório e de investigação de novos casos.

Numa primeira etapa é feito o pré-processamento dos textos para tratar os dados, utilizando a biblioteca Pandas. Ruídos como códigos HTML presentes no texto bruto são removidos e o texto é uniformizado no que tange ao uso de maiúsculas e minúsculas. Algumas outras técnicas de pré-processamento são empregadas para aumentar a eficiência das etapas seguintes.

Em seguida, são efetuadas buscas eficientes com o ElasticSearch, um motor de busca que ranqueia textos mais relevantes através de algoritmos de aprendizagem de máquina.

A partir do resultado de uma busca específica, é construída uma representação vetorial dos textos utilizando o algoritmo *text frequency - inverse document frequency*[1] e uma redução de dimensionalidade através do algoritmo *uniform manifold approximation*[2], onde o usuário visualiza em um gráfico uma representação dos textos em pontos do plano colocados de modo que pontos próximos correspondem a textos semelhantes, podendo agilmente rotular agrupamentos de textos, através de uma interface construída com a biblioteca *Dash* do *Plotly*.

Assim, posteriormente pode-se executar estratégias de classificação dos atendimentos através de algoritmos de aprendizagem supervisionada de máquina, permitindo um ganho de eficiência em buscas maiores.

O software foi hospedado nos servidores do LED-UFAL e as instâncias dele foram criadas para apoiar o trabalho de entendimento da base de textos e acessadas pelos usuários da Secretaria de Saúde através de login e senha.

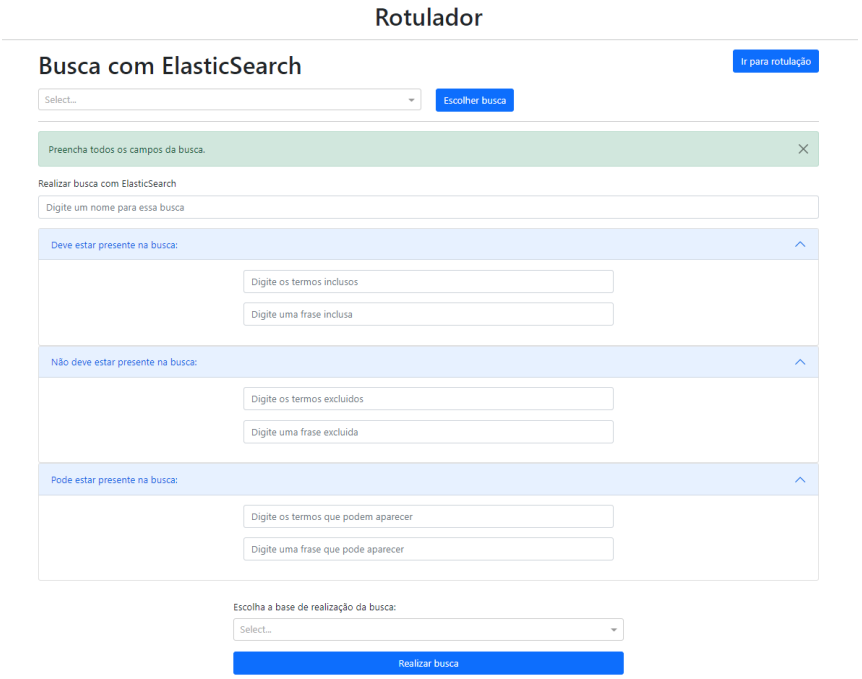



**Figura 4.** Tela inicial da plataforma, desenvolvida para analisar, visualizar e rotular textos não-estruturados de prontuários médicos.

Uma equipe de rotuladores foi montada composta por dois bolsistas do curso de graduação em Medicina da UFAL e dois funcionários da SMS de Florianópolis para realizar buscas nos prontuários de atendimentos médicos utilizando o Rotulador. O objetivo foi assinalar respostas às perguntas do CRF Post Covid de acordo com o conteúdo dos textos não-estruturados. O seguinte fluxo foi estabelecido:

[1] "Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings." https://aclanthology.org/2021.gwc-1.24. Acessado em 26 nov.. 2022.
[2] "Unsupervised Learning to Subphenotype Heart Failure Patients ...." 8 jun.. 2021, https://link.springer.com/chapter/10.1007/978-3-030-77211-6_24. Acessado em 26 nov.. 2022.

1- Inicialmente é escolhida uma pergunta de um item do CRF Post Covid e esta é atribuída a um usuário da equipe a ser respondida retrospectivamente através de dados não estruturados dos prontuários.
2- Em seguida o usuário realiza uma busca nos prontuários através do Rotulador web desenvolvido pelo LED, utilizando palavras-chave relacionadas ao item.
3- Com base na vetorização automática dos textos, uma nova tela do Rotulador é mostrada (vide imagem abaixo) onde uma clusterização de atendimentos é exibida para ser rotulada.
4- O usuário seleciona um ponto (ou conjunto de pontos), lê o conteúdo do texto e atribui uma resposta ao item analisado.
5- O objetivo é rotular o máximo de pontos possíveis para gerar uma base de prontuários rotulados com supervisão humana para em seguida desenvolver um algoritmo de Aprendizagem de Máquina que possa rotular todos os prontuários restantes.


**Figura 5.** Tela seguinte da plataforma para classificar prontuários médicos.

Dada a extensão do formulário e limitação do tempo, escolhemos algumas questões do módulo 2 e módulo 3 a serem respondidas através da rotulação. Para esta escolha nos baseamos na descrição da OMS de Síndrome Pós-COVID-19 e lista de principais sintomas[3]:

The most common symptoms of post COVID-19 condition include:

- Fatigue
- Shortness of breath or difficulty breathing
- Memory, concentration or sleep problems
- Persistent cough
- Chest pain
- Trouble speaking
- Muscle aches
- Loss of smell or taste
- Depression or anxiety
- Fever

Dada o curto tempo de desenvolvimento do projeto, focamos em rotular algumas perguntas do módulo 3:

**Módulo 3**

<table>
<tr><td colspan="2">3.5 New diagnosis of illness or complication related to COVID-19</td></tr>
<tr><td colspan="2">Was the participant newly diagnosed with any illness or complication related to COVID-19 during this visit</td></tr>
<tr><td colspan="2">Cardiovascular: ☐Yes ☐No ☐Unknown If yes, please specify diagnosis from the list below:</td></tr>
<tr><td>Atrial arrhythmia: ☐Yes ☐No ☐Unknown</td><td>Ventricular arrhythmia: ☐Yes ☐No ☐Unknown</td></tr>
<tr><td>Arterial thrombosis: ☐Yes ☐No ☐Unknown</td><td>Chronic heart failure: ☐Yes ☐No ☐Unknown</td></tr>
<tr><td>Coronary aneurysms: ☐Yes ☐No ☐Unknown</td><td>Deep vein thrombosis: ☐Yes ☐No ☐Unknown</td></tr>
<tr><td>Myocarditis: ☐Yes ☐No ☐Unknown</td><td>Pericarditis: ☐Yes ☐No ☐Unknown</td></tr>
<tr><td colspan="2">Neurological: ☐Yes ☐No ☐Unknown If yes, please specify diagnosis from the list below:</td></tr>
</table>

[3] "Post COVID-19 condition - World Health Organization (WHO)." 16 dez.. 2021, https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition. Acessado em 24 nov.. 2022.

## 5.2 - Classificação de textos

Considerando como ponto de partida o resultado da tarefa de busca, visualização e rotulação descrita anteriormente, cujo resultado é um conjunto de textos *rotulados* de acordo com um conjunto de rótulos, foram implementados três algoritmos de aprendizagem de máquina supervisionada para a tarefa de classificação de textos em Saúde[4]:

- K-nearest neighbors (KNN)[5];
- Support Vector Machine (SVM)[6];
- Fasttext[7].

Cada um desses algoritmos se propõe a classificar um dado texto de atendimento em categorias específicas, determinadas pela condição que se deseja investigar. Isso permite responder estatisticamente uma série de questões do formulário post covid-19 CRF com os dados retrospectivos, bem como realizar um monitoramento periódico para detectar novos casos e fazer acompanhamento prospectivo.

Foram feitos experimentos para 4 tipos de entrada diferentes para os classificadores:

- O texto completo como entrada;
- Um n-grama (lista com n palavras) em torno de uma palavra de maior importância para promover atenção ao contexto em volta da palavra de maior importância e evitar ruídos. Aqui n foi escolhido como 10, 12 e 14.

Seguem um exemplos ilustrativos:

Texto completo:

Descrição do prontuário: registrado por: gabriel regueira dutra medicamentos receitados - receita básica (p) soro fisiológico 0,9% - 500 ml - lavar o nariz 3 vezes ao dia - 1 frasco(s) Evolução do paciente: igdum 27+0 (incerta) igusg (12+ 0) 25+1 g1 comorbidades: arritmia sinusal medicaes em uso: nega alergias: nega s - queixa-se de dor abdomial associado a endurecimento uterino, com durao de horas, at 1 dia todo, com irradiao da dor para baixo ventre e dorso. episdios recorrentes, no ritmados. h 2 meses com epistaxe diria. nega sangramento em cavidade oral, nas fezes ou vaginal. nega alteraes de hbito intestinal. nega disuria. queixa de leso hipocrmica em brao direito. nega febre. refere aumento do corrimento, amarelado, sem cheiro, com grumos e prurido eventual. nega perda lquida. boa mf. o - beg, corada, hidratada ac: rr, 2t, s/ sopros, fc 112 bpm pa 110/70 p: 57,8 kg afu 23cm mf+ bcf 144 du ausente mmii sem edema a - gestante 2 trimestre - ganho excessivo de peso dor abdomnial sem sinais de alarme epsitaxe sem sinais de alarme p - oriento e

[4] "AI in health and medicine - Nature." 20 jan.. 2022, https://www.nature.com/articles/s41591-021-01614-0. Acessado em 26 nov.. 2022.
[5] "Large scale biomedical texts classification: a kNN and an ESA ...." 16 jun.. 2016, https://link.springer.com/article/10.1186/s13326-016-0073-1. Acessado em 26 nov.. 2022.
[6] "N-gram support vector machines for scalable procedure and ...." https://academic.oup.com/jamia/article/21/5/871/761055. Acessado em 26 nov.. 2022.
[7] "Medical-Based Text Classification Using FastText Features and ...." 31 ago.. 2021, https://link.springer.com/chapter/10.1007/978-3-030-86472-9_15. Acessado em 26 nov.. 2022.

tranquilizo, oriento alimentao e ganho de peso. prescrevo sf para lavagem nasal fazer vacinas retorno 23/09 Atendimento com especialista: Investigação de agravo: Receituario: SORO FISIOLÓGICO 0,9% - 500 ML - Posologia: Lavar o nariz 3 vezes ao dia

10-grama do texto:

0 25 1 g1 comorbidades arritmia sinusal medicaes em uso nega

Para cada escolha de tipo de entrada, vetorizador e classificador, foram realizadas várias repetições da seleção de textos para treinamento e textos para teste, para aumentar a robustez do processo e evitar viés de escolha do conjunto de treino/teste. Além disso, foram feitas análises no balanceamento entre classes, considerando duas situações: classes balanceadas (50% classe 1/50% classe 0) e classes desbalanceadas, seguindo aproximadamente a distribuição natural dos textos rotulados (75% classe 1/25% classe 0).

Assim, o total de análises foi

<table>
<tr><th>Número de opções de entrada</th><th>Número de opções de vetorizadores + classificadores</th><th>Opções de balanceamento entre classes</th><th>Total de setups de classificadores</th></tr>
<tr><td></td><td>3-NN</td><td rowspan="3">75%/25% - treino<br>75%/25% - teste</td><td rowspan="6">48</td></tr>
<tr><td>texto completo</td><td>5-NN</td></tr>
<tr><td>sequência de 10 palavras</td><td>7-NN</td></tr>
<tr><td>sequência de 12 palavras</td><td>SVC</td><td rowspan="3">50%/50% - treino<br>75%/25% - teste</td></tr>
<tr><td>sequência de 14 palavras</td><td>SVC</td></tr>
<tr><td></td><td>Fasttext</td></tr>
</table>

Por exemplo, um dos 48 setups de experimento possíveis seria:

Entrada: frase com 10 palavras em torno da palavra arritmia
Classificador: TFIDF+KNN (K=3)
Proporção entre as classes para treino: 50%/50%
Proporção entre as classes para treino: 75%/25%
Número de simulações independentes: 100

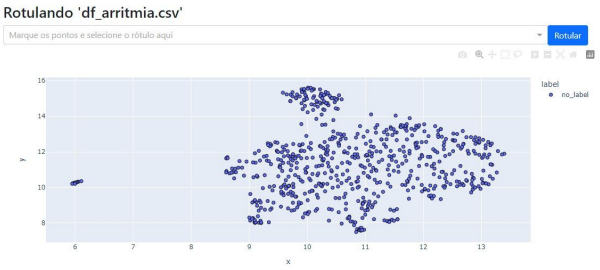
Como ilustração, iremos descrever os resultados obtidos através dessa metodologia na mineração de dados de itens da Seção 3.5: *“new diagnosis of illness or complication related to COVID-19”* e do processo de validação.

**Questão principal:** determinar novos casos de arritmia atrial ou ventricular.

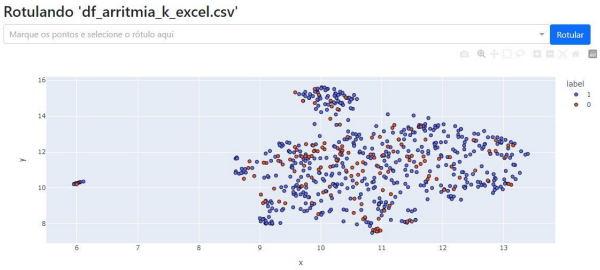
Foram realizadas as seguintes etapas:

Etapa 1: Busca pela pelos textos que tenham possíveis casos que relatam arritmia usando palavras-chave e visualização dos mesmos. Resultado: 653 textos

Etapa 2: Rotulação dos textos onde o paciente tem arritmia (rótulo 1) e textos onde o paciente não tem arritmia (rótulo 0). Resultado: 517 textos com rótulo 1 e 136 textos com rótulo 0.




VISUALIZAÇÃO SEM RÓTULOS




VISUALIZAÇÃO COM RÓTULOS

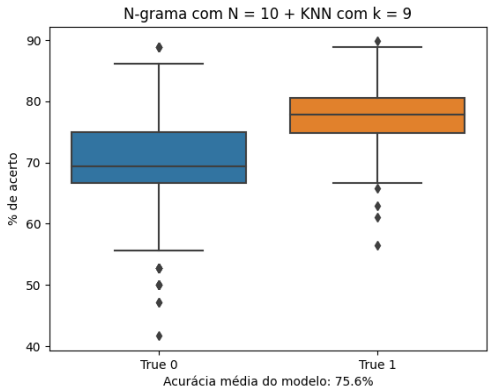
Etapa 3: Construção de classificadores para classificar os textos. As classes são nomeadas com números 1, significando que o texto informa que o paciente tem arritmia e 0, significando que o texto não informa que o paciente tem arritmia.

Entrada: frase com 10 palavras em torno da palavra arritmia
Classificador: TFIDF+KNN (K=3)
Proporção entre as classes para treino: 110/100
Proporção entre as classes para treino: 108/36
Número de simulações independentes: 1000

Etapa 4 - Validação cruzada dos classificadores.

Conforme já mencionamos, para cada escolha de tipo de entrada, vetorizador e classificador, foram realizadas 1000 repetições da seleção de textos para treinamento e textos para teste, para aumentar a robustez do processo e evitar viés de escolha do conjunto de treino/teste. Além disso, foram feitas análises no balanceamento entre classes, considerando duas situações: classes balanceadas (50% classe 1/50% classe 0) e classes desbalanceadas, seguindo aproximadamente a distribuição natural dos textos rotulados (75% classe 1/25% classe 0). A que obteve melhor desempenho segundo essa estratégia

foi o classificador KNN com K=9 e entrada um n-grama de tamanho. Seguem os resultados do modelo:




# 6 - Principais limitações encontradas e propostas de estratégias complementares

Considerando o estudo realizado, enumeramos a seguir as principais dificuldades e limitações encontradas no preenchimento do formulário:

1. **Acesso à base de dados do município:** houve uma limitação burocrática no acesso, ligada à LGPD, que atrasou o início da análise dos dados. Tal limitação foi mitigada através de um convênio entre o LED/UFAL e o município.
2. **Falta de documentação da complexa base de dados do CELK**: A extração dos dados foi excessivamente complicada por falta de documentação e suporte dos desenvolvedores da plataforma CELK: não houve canal de comunicação entre desenvolvedores e nossa equipe para questionamento ou mapeamento das perguntas na plataforma de preenchimento para as respostas armazenadas nas bases de dados relacionais do *back-end*. As bases de dados relacionais eram compostas de centenas de tabelas, as quais não foram documentadas e várias das quais eram obsoletas, compondo a dificuldade na identificação de dados pertinentes. Além disso, algumas informações críticas, como identificadores dos pacientes e atendimentos, não batiam entre as mostradas na interface e as armazenadas nas bases do *back-end*; isto causou grandes problemas na validação dos trabalho até que fosse descoberto.
3. **Intermitência do acesso aos servidores do CELK:** o trabalho foi frequentemente interrompido por conta de perda de acesso periódica às bases do CELK: quase que

todas as semanas, nossas permissões de acesso eram automaticamente invalidadas e era necessário que a Secretaria de Saúde abrisse chamado junto à empresa responsável para resolução, o que custava de um a dois dias no meio da semana de trabalho de parte da equipe.

4. **Distinção entre consultas médicas e consultas de outras especialidades e serviços:** Não conseguimos distinguir no backend consultas médicas de outros tipos de serviços, como atendimento psicológico e retirada de exames;
5. **Adaptação para uma análise retrospectiva de um estudo inicialmente proposto para ser prospectivo**: alguns itens do formulário não podem ser respondidos retrospectivamente.
6. **Baixo preenchimento de algumas informações estruturadas:** como era esperado em um serviço ambulatorial, houve um baixo preenchimento de alguns dados estruturados, limitando os resultados obtidos pela mineração dos dados.
7. **Tempo curto para desenvolvimento do Processamento de Linguagem Natural:** considerando a natureza dos textos, é necessário mais tempo para implementação e validação em maior escala da estratégia de via processamento de linguagem natural.

Considerando a complexidade do problema, houve um substancial avanço durante os três meses de sua execução e acreditamos que há um grande potencial de evolução. O trabalho realizado junto com a secretaria de saúde promoveu um avanço no entendimento do problema e suas dificuldades. A seguir, listamos as principais propostas para mitigar as dificuldades encontradas:

1. **Desenvolvimento de pesquisas, algoritmos e interfaces de processamento de linguagem natural para textos médicos:** tal estratégia permite a construção de classificadores eficientes de texto que podem ajudar na mineração eficiente dos dados, recuperação de informação e realizar o monitoramento ativo de novos casos, permitindo a eventualmente estudos prospectivos sem tanto custo humano. Foram construidos protótipos nesse estudo para avaliar o potencial de tal estratégia, que podem ser úteis numa implementação e projeto posterior.

2. **Desenvolvimento de interfaces computacionais para otimizar o trabalho da vigilância epidemiológica:** considerando que boa parte dos algoritmos de processamento natural requerem um trabalho custoso de rotulação, recomenda-se que haja a construção de interfaces que incorporem

naturalmente a inteligência produzida no trabalho diário dos profissionais de saúde, de modo que possam ser consumidas por tais algoritmos;

3. **Espelhamento periódico da base de dados em uma estrutura computacional própria do município:** isso permitiria o acesso a base de dados por parte dos servidores do município e diminuiria a indisponibilidade.

4. **Documentação da base de dados independentemente do software de gestão de saúde utilizado:** isso permitiria o acompanhamento periódico e mapeamento dinâmico das mudanças realizadas pela empresa contratada para a gestão dos dados de saúde.

# 7 - Conclusão e alimentação da plataforma global

Análise de bases massivas de dados de saúde apresenta-se como elemento fundamental no estudo de doenças emergentes, suas complicações e desenvolvimento célere de políticas de saúde pública adequadas para preservação de nossas comunidades. Frente a este desafio, o presente projeto propôs-se dois desafios na utilização da base de dados ambulatoriais do Município de Florianópolis:

- Mineração e automação de preenchimento do máximo de perguntas [do formulário Post COVID-19 CRF] possíveis com base em dados estruturados da base municipal de saúde;

- Desenvolvimento de protótipo de ferramentas e pipeline abertos para que a Secretaria de Saúde de Florianópolis e quaisquer outra entidade de saúde pudesse analisar textos livres e não estruturados de suas bases para extração de inteligência epidemiológica.

O presente relatório resume os desenvolvimentos nestas duas frentes.

Na frente de preenchimento do formulário Post COVID-19 CRF, alinhados ao procedimento prospectivo adotado pelos parceiros brasileiros do projeto, identificamos 28.817 pacientes de interesse ao estudo. Por meio dos campos estruturados da base CELK, foram preenchidos 42 campos do formulário Post COVID-19 CRF na forma de uma base entregue juntamente a este relatório.

Estes campos e seus mapeamentos na base CELK do município de Florianópolis são identificados na **Seção 3**, enquanto que os critérios de filtragem para identificação de pacientes de interesse e estatísticas resumidas são descritos na **Seção 3**.

Referente ao desenvolvimento de pipeline para mineração de prontuários não estruturados para compilação de inteligência epidemiológica, entregamos protótipo de rotulador para classificação de conjunto de treinamento de prontuários para atender a perguntas específicas. Este conjunto rotulado de prontuários é então utilizado para o treinamento de redes neurais para classificação de bases massivas de prontuários. Exploramos a aplicação do protótipo para rotulação de conjuntos de prontuários para buscar manifestação dos 10 principais sintomas associados à síndrome pós-covid. Exploramos variações de vetorizadores e de classificadores de texto para classificação de prontuários, visando construir uma metodologia para preenchimento das questões do questionário e investigação de novos casos. Tal metodologia foi testada para identificar referências à anosmia e à arritmias cardíacas, como ilustrado na **Seção 5**. As limitações e propostas de mitigação das dificuldades encontradas no processo são discutidas na **Seção 6**.

Os produtos deste estudo correspondem às limitações de acesso de dados, à escala e complexidade dos mesmos, e às limitações operacionais e programáticas. Salientamos a relevância do desenvolvimento de pipelines para obtenção de inteligência epidemiológica no enfrentamento de desafios sanitários modernos. Acreditamos que este projeto foi um primeiro passo importante, e buscaremos o aprimoramento contínuo destas ferramentas.

# REFERÊNCIAS

[1] Plataforma Clínica Global da OMS para COVID-19. Dados para a resposta da saúde pública. Relatório sobre a caracterização clínica da COVID-19 Brasil. Junho 2021. <https://iris.paho.org/handle/10665.2/54817>

[2] Anschau, Fernando, et al. "Cohort study protocol of the Brazilian collaborative research network on COVID-19: strengthening WHO global data." BMJ open 12.11 (2022): e062169.

[3] Bowe, Benjamin, Yan Xie, and Ziyad Al-Aly. "Acute and postacute sequelae associated with SARS-CoV-2 reinfection." Nature Medicine (2022): 1-8.

[4]Greenhalgh, Trisha, et al. "Long covid—an update for primary care." bmj 378 (2022).

[5] World Health Organization. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021. No. WHO/2019-nCoV/Post_COVID-19_condition/Clinical_case_definition/2021.1. World Health Organization, 2021.

[6] de Miranda, Daniel AP, et al. "Long COVID-19 syndrome: a 14-months longitudinal study during the two first epidemic peaks in Southeast Brazil." Transactions of The Royal Society of Tropical Medicine and Hygiene (2022).

[7] "AI in health and medicine - Nature." 20 jan.. 2022, https://www.nature.com/articles/s41591-021-01614-0..

[8] "Medical-Based Text Classification Using FastText Features and ...." 31 ago.. 2021, https://link.springer.com/chapter/10.1007/978-3-030-86472-9_15

# ANEXOS

## Lista de arquivos que acompanham o relatório:

(disponíveis no link: https://github.com/goedert/LED_Floripa_Covid_OMS-Resultados.)